应用昆虫学报 *Chinese Journal of Applied Entomology* 2013,50(1):288 - 292. DOI:10.7679/j.issn.2095 - 1353.2013.038

昆虫系统发育重建的常用方法及步骤*

秦 洁** 张爱兵***

(首都师范大学生命科学学院 北京 100048)

摘 要 本文介绍了昆虫系统发育重建研究的步骤,常用方法及相关软件的使用,指出了不同研究方法的优缺点及适用范围,分析了系统发育重建存在的问题,从而为相关研究的开展提供了参考。

关键词 系统发育重建,系统发育树,序列比对,昆虫系统发育

Commonly used methods for phylogenetic reconstruction of insects

QIN Jie** ZHANG Ai-Bing***

(College of Life Sciences, Capital Normal University, Beijing 100048, China)

Abstract We briefly introduce the commonly used methods and programs for phylogenetic reconstruction of insects. We discuss the advantages and disadvantages of these methods and evaluate them with an empirical study.

Key words phylogenetic reconstruction, phylogenetic tree, sequence alignment, phylogeny of insects

对于解决物种之间的关系以及追溯生物界不同生物类型的起源和进化等问题,系统发育重建是很重要的解决方法,并且受到日益关注。随着测序技术的发展,越来越多的系统发育分析方法和软件层出不穷,系统发育分析的研究也到达了一个新的高度,几乎已经渗透到生物学的所有分支,包括动物学、植物学、生态学等。昆虫作为动物界中最庞大类群之一,目前已知 100 万种以上,自古以来就与人类生活、健康和经济密切相关,因此,进一步探究昆虫的系统发育关系非常重要。大量实例研究表明,系统发育研究在昆虫中的应用非常广泛,且发展较快。

昆虫系统发育重建可以基于多种信息特征来进行研究,比如形态学特征、解剖学特征和发育学特征等,本文主要介绍基于分子序列信息进行系统发育重建研究的步骤、常用方法及软件的使用。

1 方法

昆虫系统发育重建研究主要包括 4 个步骤:序列的获得、校对与编辑、序列比对和构建系统发育树,如图 1 所示。

1.1 序列的获得

进行系统发育分析,首先要获得研究对象的分子序列(核苷酸序列或氨基酸序列),这些序列可以由实验获得,也可以从相关的数据库(例如 Genebank,EMBL,DDBJ 等)下载得到。

1.2 序列校对与编辑

序列的校对与编辑主要是针对由实验得到的核苷酸序列。通过实验得到核苷酸序列后,首先需要检查其测序的峰图,逐一寻找是否有误读的碱基及质量较差的峰,然后对其进行手动校正或者删除。这一步常用软件包括 chromas 和 BioEdit Sequence Alignment Editor,均可以直接打开 DNA 测序结果(.ABI 格式的文件)来观察测序峰图。一般来说,一段序列的前后 20 bp 信号峰图比较杂乱,要进行删除,中间部分出现酒精峰和杂带(图 1)时也要进行校正或重测。

1.3 序列比对

进行系统发育重建的关键步骤是根据所要分

* 资助项目:国家自然科学基金项目(31272340);北京市人才强教项目(PHR201107120)。

** E-mail:qinjie0912@163.com

*** 通讯作者,E-mail:zhangab2008@gmail.com

收稿日期:2012-12-21,接受日期:2012-12-30

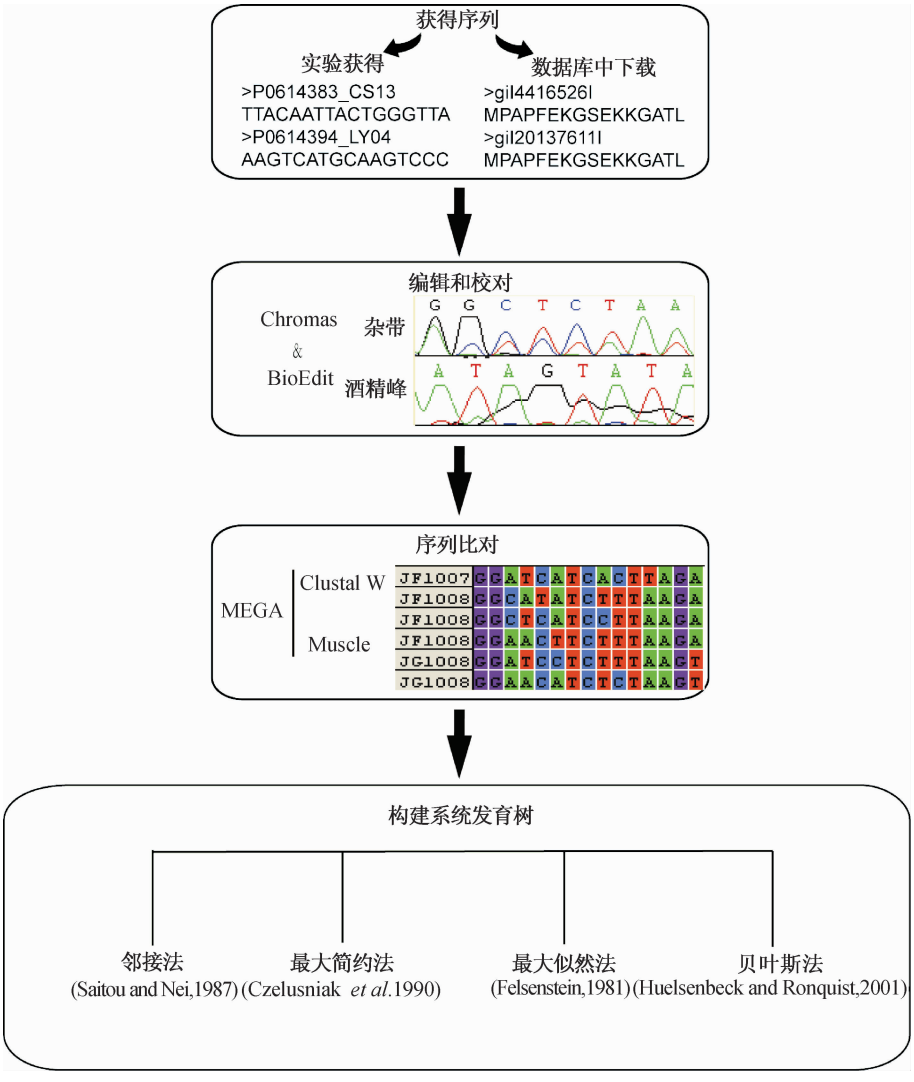

图 1　昆虫系统发育重建的主要步骤

Fig. 1　The main steps of insect phylogenetic reconstruction

析序列的同源性进行比对。对于一些亲缘关系比较远的类群,这能够降低因为空位(gap)的存在所造成的差异。常用的序列比对软件有 Clustal W (Thompson *et al.*,1994)和 Muscle(Edgar,2004),其中 Clustal W 是目前较为主流的比对软件。此外,MEGA5(Tamura *et al.*,2011)这一软件整合了以上两种比对软件,并且提供了图形化的界面,使用起来也较为方便。

1.4　构建系统发育树

昆虫的系统发育研究主要是对构建的系统发育树进行分析,所谓系统发育树是由进化分支(branch)连接各结点(node)组成的。

系统发育关系需要从序列或其他数据中推断,并不能够直接观察得到。系统发育重建的方法有两类,基于距离或者是基于特征(Yang and Rannala,2012)。基于距离的方法是计算序列两两之间遗传距离,并用所得到的距离矩阵进行系统发育树的构建。基于特征的方法是比较排列中的所有序列,考虑到每个特征(对齐中的每个位点)计算每棵系统发育树的得分。一般来说,生物学中系统发育树的构建方法主要有 4 种:邻接法(Saitou and Nei,1987)、最大简约法(Czelusniak *et al.*,1990)、最大似然法(Felsenstein,1981)和贝叶斯方法(Huelsenbeck and Ronquist,2001)。下面具体介绍一下以上 4 种方法的原理和相关软件的使用。

1.4.1　邻接法(Neighbor-Joining,NJ)　邻接法

是基于距离的方法,它根据一定的进化模型,通过比较推导出各个序列之间的进化距离,构建一个距离矩阵,再依据此矩阵和一定的算法,依次将序列合并聚类,构建系统发育树。

基于距离的方法(尤其是邻接法)其一大优势是它的计算效率很高,应用的聚类算法比较快,且不需要像简约法和最大似然法一样在一个最优的标准下比较许多的树。它只要基于一个符合实际的进化模型,再选择不同的遗传距离计算方法。但是,距离的方法对于分歧比较大的序列常常表现欠佳。因为,较大的分歧涉及到的抽样误差较大,而不被大多数距离的方法(如邻接法)所考虑(Yang and Rannala,2012);而且,分歧比较大的序列涉及到空位对准的问题,距离的方法对于序列比对中的空位很敏感,从而使计算出现问题。基于以上原因,邻接法在分析序列分歧水平低的较大数据集时比较有用。构建邻接树常用的的软件是 MEGA5/PHYLIP(Felsenstein,2005)/PAUP* 4.0(Swofford,2000)。

1.4.2 最大简约法(maximum parsimony, MP) 最大简约法是基于特征的方法,根据信息位点提供的各序列间的碱基替换情况构建所有可能的树,然后筛选出含最小替换数的最优树。

最大简约法不需要像距离法或似然法那样在处理序列关系时需要进化模型,它源于形态学的研究,比较简单,很容易描述和理解,还有助于开发有效的计算机算法。当序列分歧度较低时,无需模型的简约法可以获得比其他方法更可靠的系统发育树。但是,简约法缺乏明确的模型假设,这使它几乎不可能在构建系统发育树时利用物种进化过程中的信息。当被检验的信息位点数比较少,且存在较多的回复突变或平行突变的时候,可能会给出一个不合理或者错误的推导结果,即分支长度可能会被大大的低估。此外,当简约法为了校正相同位点的多个碱基替换时,会遇到"长枝吸引"(long-branch attraction)问题。构建简约树常用的软件是 MEGA5/ PAUP* 4.0。

1.4.3 最大似然法(maximum likelihood, ML) 最大似然法也是基于特征的方法,它根据一个特定的碱基替代模型来分析序列,使得用模型产生的数据与真实数据之间的相似程度最高(得到最大的似然值),在计算过程中,通过计算每一个拓扑结构的似然值,然后挑出其中似然值最大的拓扑结构作为最优树。

最大似然法充分有效地利用所有数据信息,其进化概率模型采用了标准的统计方法,并且所有的进化模型假设都比较明确,可以被评估和改进。最大似然法采用多样复杂的进化模型更接近生物真实的进化历史,这也是相对于最大简约法的一个主要的优势。但是,最大似然法涉及大量的计算,特别是在似然准则下搜索树时对计算能力的要求很高。此外,最大似然法对于模型的依赖性很高,如果模型使用不当,这种方法可能会有不好的统计学特性。构建最大似然树常用的的软件是 MEGA5/PhyML(Guindon and Gascuel,2003)/PAUP*4.0。

1.4.4 贝叶斯方法(Bayesian inference) 贝叶斯方法也是基于特征的方法,它基于最大后验概率原理,评估所推断的那棵树的可靠性,具有最大后验概率的那棵树最接近于真实。

贝叶斯方法和最大似然法一样,有许多共同的统计特性,如一致性和有效性,它也拥有更接近生物真实的替代模型。但是,用贝叶斯方法计算的树和分支上的后验概率常常被高估,它对于模型的选择很敏感,使用较简单的模型时,常常会导致后验概率比较高。先验概率可以合并树和参数的先验信息,然而这样的信息是很少的,基本上所有的分析都是使用电脑程序中默认的先验值。而且,一个看似普通的先验值对后验概率会有影响。例如,最近被指出,在使用 MrBayes(Huelsenbeck and Ronquist,2001)时,枝长独立先验指数对于树的长度能够产生不合理的先验值,导致在许多数据集中出现不合理的树长。因此,在进行贝叶斯分析的时候,评估先验概率对于后验概率的影响是很重要的。另外,马尔可夫链蒙特卡罗(MCMC)涉及大量的计算,在分析大型数据集时,MCMC 的收敛性和混合问题很难被发现或纠正。构建贝叶斯树常用的的软件是 MrBayes。

2 应用分析

在这里,以 MEGA5 软件构建邻接树为例,介绍昆虫系统发育重建的操作步骤。

首先,将校对后的目的序列合并为一个 fasta 文件(图 2)进行序列比对。打开 MEGA5 使用界面——File——Open a File/session——选择文件——出现的对话框中选择 Align ——选择要比

对的序列——Alignment ——选择比对方法(Alignment by Clustal W/Muscle)——弹出的界面可以更改参数,点击 OK 得到结果。将得到的序列比对结果保存,以便进行后续的构建系统发育树等操作。

然后,利用比对好的序列构建系统发育树。打开 MEGA5 使用界面——File——Open a File/session——载入比对后的文件——出现的对话框中选择 Analysis——Phylogeny——Construct/Test Neighbor-Joining Tree——在出现的 Analysis Preference 对话框中选择参数,例如 Test of Phylogeny 中选择 Bootstrap(Felsenstein,1985),No. of Bootstrap Replications 中选择数量——得到系统发育树。

```
>BHS100705061_Pangrapta_disruptalis
GTTTACTAATTCGTGCTGAATTAGGTAATCCAGGATCTTTAATTGGTGACGATCAAATTT
>BHS100705132_Callopistria_juventina
GTTTATTAATTCGTGCTGAATTGGGAAACCCAGGATCATTAATTGGAGATGATCAAATTT
>BHS100703102_Hadjina_chinensis
GATTATTAATTCGTGCTGAATTAGGAACCCCAGGATCTTTAATTGGAGATGATCAAATTT
>BHS100705035_Hypocala_subsatura
GATTATTAATTCGTGCTGAATTAGGTAACCCAGGATCATTAATTGGAGACGATCAAATTT
>BHS100703155_Anterastria_atrata
GTTTATTAATTCGTGCTGAATTAGGAAATCCTGGATCTTTAATTGGAGATGATCAAATTT
>BHS100703161_Niphonyx_segregata
GATTATTAATTCGAGCTGAATTAGGAAATCCAGGATCTTTAATTGGAGATGATCAAATTT
```

图 2　fasta 文件格式

Fig. 2　Format of the fasta file

按照以上的步骤,即可得到系统发育树,系统发育树反映了研究对象之间的系统进化关系,树分支的图像称为拓扑结构,其中分支长度表示该分枝进化过程中变化的程度,树中的每一节点表示一个分类学单元(属、种群、个体等),节点包括内节点和外节点(张树波和赖剑煌,2010),内节点代表生物分子之间的进化位置,外节点代表生物分子;树中的分支定义了分类单元(祖先与后代)之间的关系,一般一个分支只能连接两个相邻的节点。

构建最大简约树和最大似然树也可用 MEGA5 软件,但是需要注意的是构建最大似然树要选择适合的模型假设;构建贝叶斯树一般是选用 MrBayes 软件,在其软件手册中有详尽的解释,在此就不做介绍了。

3　讨论

一般采用 Bootstrap 的方法来评估所构建系统发育树的可靠性,Bootstrap 方法的一般原理是:从待分析的序列中,随机有放回的不断抽取其中一列,构成相同长度且新排列的序列,反复进行后得到多组新序列,对这些序列构建系统发育树,以此评价所构建的系统进化树的可靠性。一般情况下,Bootstrap 重复选择 500 ~ 1 000 次。

在过去的十几年里,系统发育重建研究已取得很多进展,这一方面得益于计算机技术的发展,另一方面得益于生命科学的进步。但是系统发育重建其实是研究已发生的历史,历史只能被推断或者评估,因此系统发育重建会存在很多问题,这一方面是由于生命进化过程的复杂性,仅仅从分子水平上分析其系统发育关系并不能得到可靠的结果,另一方面是因为用计算机方法分析系统发育关系仍不可避免的会出现错误。

现在的系统发育分析只是应用某些系统发育程序所得到的结果,其可靠性和实用性还存在许多争议,但是它仍然在一定程度上反应了真实的系统发育关系,而且,随着生命科学的不断发展和系统发育分析方法的不断完善,对于物种系统发育关系的了解会更加深入且更接近于生物真实的历史,相信在不久的将来,昆虫系统发育重建的研

究会取得新的突破。

参考文献(References)

Czelusniak J, Goodman M, Moncrief ND, Kehoe SM, 1990. Maximum parsimony approach to construction of evolutionary trees from aligned homologous sequences. *Methods Enzymol.*, 183:601 – 615.

Edgar RC, 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792 – 1797.

Felsenstein J, 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17(6): 368 – 376.

Felsenstein J, 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4): 783 – 791.

Felsenstein J, 2005. Phylip: Phylogenetic Inference Program, Version 3.6. (Univ. of Washington, Seattle).

Guindon S, Gascuel O, 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, 52(5):696 – 704.

Huelsenbeck JP, Ronquist F, 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754 – 755.

Saitou N, Nei M, 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406 – 425.

Swofford DL, 2000. PAUP*: Phylogenetic Analysis by Parsimony (and Other Methods) 4.0 Beta (Sinauer Associates, Massachusetts).

Tamura K, Peterson D, Peterson R, Stecher G, Nei M, Kumar S, 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.*, 28: 2731 – 2739.

Thompson JD, Higgins DG, Gibson T, 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22(22):4673 – 4680.

Yang Z, Rannala B, 2012. Molecular phylogenetics: principles and practice. *Nature Review*, 13:303 – 314.

张树波, 赖剑煌, 2010. 分子系统发育分析的生物信息学方法. 计算机科学, 37(8):47 – 51.